IPSJ Transactions on Bioinformatics Vol.6 35–42 (July 2013)

[DOI: 10.2197/ipsjtbio.6.35]

Original Paper

SCPSSMpred: A General Sequence-based Method for Ligand-binding Site Prediction

Chun Fang[1,3,a)] Tamotsu Noguchi[2,3,b)] Hayato Yamana[1,c)]

Received: February 11, 2013, Accepted: April 23, 2013, Released: July 10, 2013

Abstract: In this paper, we propose a novel method, named SCPSSMpred (Smoothed and Condensed PSSM based prediction), which uses a simplified position-specific scoring matrix (PSSM) for predicting ligand-binding sites. Although the simplified PSSM has only ten dimensions, it combines abundant features, such as amino acid arrangement, information of neighboring residues, physicochemical properties, and evolutionary information. Our method employs no predicted results from other classifiers as input, i.e., all features used in this method are extracted from the sequences only. Three ligands (FAD, NAD and ATP) were used to verify the versatility of our method, and three alternative traditional methods were also analyzed for comparison. All the methods were tested at both the residue level and the protein sequence level. Experimental results showed that the SCPSSMpred method achieved the best performance besides reducing 50% of redundant features in PSSM. In addition, it showed a remarkable adaptability in dealing with unbalanced data compared to other methods when tested on the protein sequence level. This study not only demonstrates the importance of reducing redundant features in PSSM, but also identifies sequence-derived hallmarks of ligand-binding sites, such that both the arrangements and physicochemical properties of neighboring residues significantly impact ligand-binding behavior[*1].

Keywords: sequence-based, ligand-binding, prediction, simplified PSSM

1. Introduction

Identifying ligand-binding sites is a key step in annotating protein function and applying the knowledge in drug design. Physical experimental methods for identifying binding sites are expensive and time consuming, which makes computational methods indispensable for guiding the physical experimental analysis. Now, a variety of sequence-based tools for predicting protein's function sites in proteins exist [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14]. Previous methods are mainly based on the features of amino acid arrangement, predicted disorder probabilities, predicted solvent accessibility, predicted secondary structure, physicochemical properties, evolutionary information, and so on. The design of these predictors is complex and their performance is affected to a significant degree by other predictors. Indeed, many of these methods provided no Web services owing to the complexity of their design.

Evolutionary information included in PSSM has been considered the most effective feature for ligand-binding prediction. Raghava's group has used PSSMs for predicting many kinds of functional sites [7], [8], [9], [10], [11]. John et al. [12] have shown that conservation features are highly predictive in identifying ligand-binding sites and catalytic sites compared to the identification of other functional sites, such as protein-protein interfaces. To date, however, all these methods have used the direct output of PSSM for prediction, without considering its redundant features. While the physicochemical properties of residues have been used for ATP-binding prediction [11], they have been found to perform less effective when used alone. If physicochemical properties or structure characteristics are used in tandem with PSSM, the feature dimensions would increase significantly, and theoretically result in a higher dimensional feature space.

While many machine learning methods have been used for identifying functional sites, such as SVM, ensemble SVM, random forest, naïve Bayes, and neural network, they all encounter the same problem that, the available samples are limited due to the limitation of physics experiments, and the dimensionality of PSSM is high (20 dimensions). High-dimensional feature space invariably requires a larger number of training samples, and readily leads to over-fitting to noise data. In the case that a limited number of samples are available, if the dimensions of PSSM are not reduced, it is impossible to significantly improve the performance of predictor regardless of the machine learning method. Therefore, to identify function sites in proteins, it is necessary to find a more effective feature-encoding method that combines more predictive features into fewer dimensions.

It has been observed [13] that nucleotide-binding sites tend to occur in clusters, and that nucleotide-binding residues are gener-

[1] School of Fundamental Science and Engineering, Waseda University, Shinjuku, Tokyo 169–8555, Japan

[2] Pharmaceutical Education Research Center, Meiji Pharmaceutical University, Kiyose, Tokyo 204–8588, Japan

[3] Computational Biology Research Center (CBRC), Koto, Tokyo 135–0064, Japan

[a)] fangchun@yama.info.waseda.ac.jp

[b)] noguchit@my-pharm.ac.jp

[c)] yamana@yama.info.waseda.ac.jp

[*1] Availability: http://webapp.yama.info.waseda.ac.jp/fang/ligand2.php.

© 2013 Information Processing Society of Japan

ally conserved and flanked by less conserved residues in the sequence [14]. We have calculated the number of continuous binding residues in our datasets, and found that nearly 70% of the ligand-binding residues are continuous. Inspired by a smoothing method developed for image processing [15], Chen et al. [16] have shown that a smoothing method incorporating the dependency on surrounding neighbors of a central residue can improve the performance in the prediction of RNA-binding sites. In our previous studies [17], we have shown that combinatorial features of PSSM with physicochemical properties outperformed combinatorial features of PSSM with structural information (e.g., solvent accessibility) in ligand-binding prediction. Furthermore, we found that using the physicochemical properties of amino acids to condense a standard PSSM can reduce redundant features and improve prediction performance [18], [19].

In this paper, we propose a novel sequence-based prediction method, named as SCPSSMpred, which uses a smoothed and condensed PSSM for ligand-binding prediction. This simplified PSSM has only ten dimensions, which is half the number of a traditional PSSM (20 dimensions). No other predicted results were used as input, i.e., all features used in this method were extracted from sequences only. For comparison purposes, three other traditional methods were also analyzed, all of them used the support vector machines (SVM). Finally, to verify its versatility, the SCPSSMpred method was tested on three classes of ligand-binding proteins.

2. Methods

2.1 Data Sets

Raghava's group has used PSSMs for many kinds of ligand-binding prediction, such as ATP, ADP, GTP, NAD, and FAD [7], [8], [9], [10], [11]. In this study, three kinds of ligands are chosen as representatives for analyzing. They are Nicotinamide adenine dinucleotide (NAD), Flavin adenine dinucleotide (FAD), and Adenosine-5'-triphosphate (ATP). Data sets were collected respectively as follows: Firstly, PDB IDs that have contact with NAD, FAD, and ATP were extracted from Supersite [20]. Next, Ligand Protein Contact (LPC) [21] was used to identify side chains containing the NAD, FAD, and ATP binding sites. Finally, redundant chains of less than 50 residues in length or with a sequence similarity > 40% were discarded. In this manner, we obtained the following three datasets:

Dataset NAD204: It includes 204 NAD-binding protein chains, containing 5,165 NIRs (NAD interacting residues) and 65,605 non-NIRs (non-NAD interacting residues), named NAD204.

Dataset FAD191: It includes 191 FAD-binding protein chains, containing 5,662 FIRs (FAD interacting residues) and 73,680 non-FIRs (non-FAD interacting residues), named FAD191.

Dataset ATP200: It includes 200 ATP-binding protein chains, containing 3,595 AIRs (ATP interacting residues) and 71,514 non-AIRs (non-ATP interacting residues), named ATP200.

Because they have different functions, the three ligands cannot be mixed for prediction. Accordingly, their predictors were designed individually, but based on the same method. The negative samples were selected randomly with an equal number of positive samples.

2.2 Continuous Binding Residue Analysis

Chen et al. [14] found that nucleotide-binding residues are usually clustered close together in protein sequences. We first calculated the continuous binding residues in the three classes of ligand-binding proteins (**Fig. 1**), and found that, the lengths of continuous binding sites are between 1 to 6 residues.

2.3 Composition of Residues Analysis

Next, the percentage differences between binding and non-binding residues (percentage of binding sites - percentage of non-binding sites) for the 20 amino acids in the three datasets were calculated. As shown in **Fig. 2**, amino acids G, H, I, F, and T — most of which are hydrophobic and aromatic amino acids — were overrepresented in NAD-binding sites. Similarly, A, R, Q, E and L were overrepresented in FAD-binding sites, and G, S, W and Y were overrepresented in ATP-binding sites.

2.4 Preference of Physicochemical Properties

The differences between binding and non-binding residues with respect to ten physicochemical properties were also calculated (percentage of binding sites – percentage of non-binding sites). **Figure 3** demonstrates that physicochemical properties such as hydrophobic, small, tiny, and aromatic have relatively higher proportions in NIRs and FIRs than non-NIRs and non-

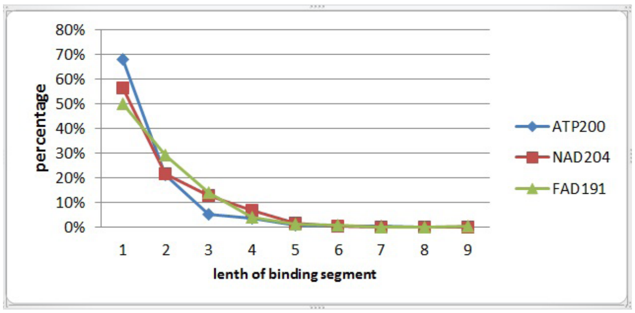


Fig. 1 Statistics of continuous binding residues.

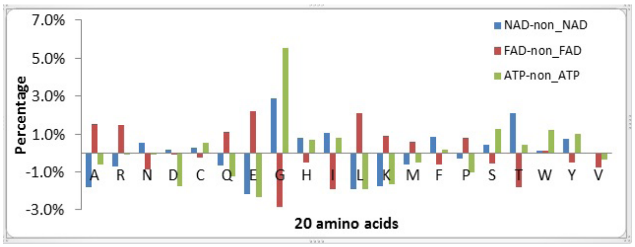


Fig. 2 Composition difference on 20 amino acids.

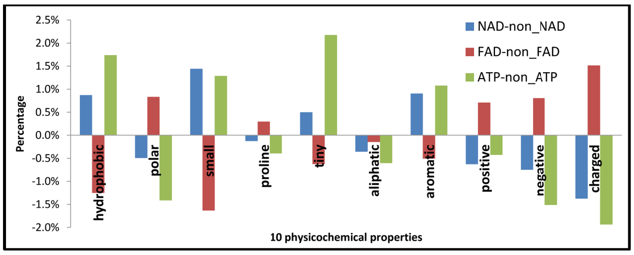


Fig. 3 Composition difference on ten physicochemical properties.

© 2013 Information Processing Society of Japan

FIRs, while physicochemical properties such as polar, positive, negative, and charged were overrepresented in AIRs compared to non-AIRs. Thus, physicochemical properties can be used as a significant feature to identify ligand-binding sites.

2.5 Prediction Model

By integrating all the above analysis, we designed our prediction model based on a smoothed and condensed PSSM which includes the information of amino acid position, arrangement of neighboring residues and physicochemical characteristics, in addition to evolutionary information. The prediction model is shown in **Fig. 4**. The detailed description of each part is explained later.

2.6 Physicochemical Features

Ten discriminative physicochemical features of residues are considered in our study. These features include hydrophobic, polar, small, proline, tiny, aliphatic, aromatic, positive, negative, and charged. Each amino acid is represented by a vector length of 10 (e.g., Ala by 1 0 1 0 1 0 0 0 0 0).

2.7 Evolutionary Information (PSSM)

Evolutionary information can be obtained from PSSMs, generated by PSI-BLAST [22] searching against NCBI non-redundant (nr) database [23] through three iterations, with an e-value of 0.001. The evolutionary information for each amino acid is encapsulated in a vector of 20 dimensions, where the size of PSSM matrix of a protein with N residues is $20 * N$. Here, 20 is the number of the standard amino acids, and N is the length of the protein.

2.8 Smoothing the Standard PSSM

Every value in a standard PSSM is calculated based on the assumption that the position of each value in the matrix is independent of the others. However, the statistics in Fig. 1 illustrate that 70% of ligand-binding residues appear continuously, indicating that binding sites are largely influenced by their neighboring residues. In order to incorporate the dependency on surrounding neighbors of a central residue, we adopt the previously published smoothing method [16], which is based upon consideration of adjacent pixels used in the spatial domain method in the field of image processing [15].

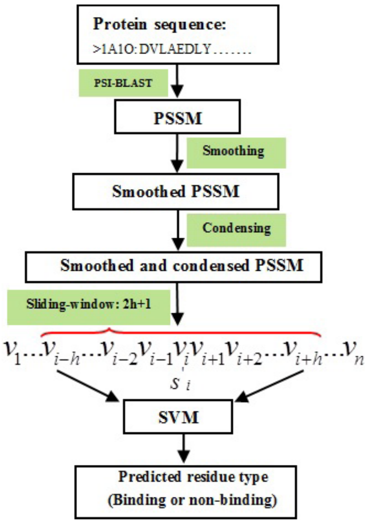
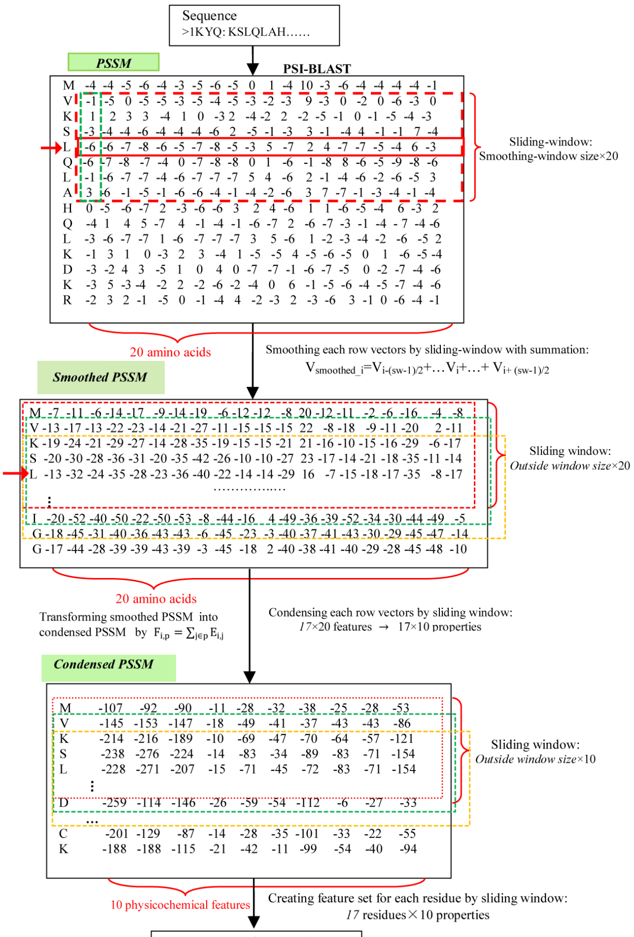


Fig. 4 Prediction model. The length of sliding window is represented by $2h + 1$, where n is the length of sequence, and v_i represent the corresponding amino acid i in the feature vector.

Firstly, in order to deal with the N-terminal and C-terminal of a protein sequence, $(sw - 1)/2$ ZERO vectors are appended to the head and tail of a standard PSSM profile, where sw is the size of a smoothing sliding window. The smoothing sliding window is then used to incorporate the evolutionary information from upstream and downstream residues. Each row vector of an amino acid residue S_i is smoothed by the summation of sw (sw is an odd number) surrounding row vectors; $V_{smoothed\_i} = V_{i-(sw-1)/2} + ...V_i + ... + V_{i+(sw-1)/2}$. **Figure 5** illustrates an example of a smoothed PSSM profile. For amino acid 'L', the first column of the vector is smoothed by the summation of $[(-1) + 1 + (-3) + (-6) + (-6) + (-1) + 3 = (-13)]$.

2.9 Condensing the Smoothed PSSM

After smoothing the standard PSSM, the predictor selects ten discriminative physicochemical properties (hydrophobic, polar, small, proline, tiny, aliphatic, aromatic, positive, negative, and charged) of a residue to condense the output of smoothed PSSM. The smoothed PSSMs are then divided into sliding windows of size m. Each window is a matrix $E_{ij}\{i = 1, ..., m, j = 1, ...20\}$,


Sequence
>1KYQ: KSLQLAH……
PSSM
PSI-BLAST
Sliding-window:
Smoothing-window size×20
20 amino acids
Smoothing each row vectors by sliding-window with summation:
$V_{smoothed\_i}=V_{i-(sw-1)/2}+\ldots V_i+\ldots+V_{i+(sw-1)/2}$
Smoothed PSSM
Sliding window:
Outside window size×20
20 amino acids
Transforming smoothed PSSM into condensed PSSM by $F_{i,p} = \sum_{j \in p} E_{i,j}$
Condensing each row vectors by sliding window:
17×20 features → 17×10 properties
Condensed PSSM
Sliding window:
Outside window size×10
10 physicochemical features
Creating feature set for each residue by sliding window:
17 residues×10 properties
Feature set for predictor

M	-4	-4	-5	-6	-4	-3	-5	-6	-5	0	1	-4	10	-3	-6	-4	-4	-4	-4	-1
V	-1	-5	0	-5	-5	-3	-5	-4	-5	-3	-2	-3	9	-3	0	-2	0	-6	-3	0
K	1	2	3	3	-4	1	0	-3	2	-4	-2	2	-2	-5	-1	0	-1	-5	-4	-3
S	-3	-4	-4	-6	-4	-4	-4	-6	2	-5	-1	-3	3	-1	-4	4	-1	-1	7	-4
L	-6	-6	-7	-8	-6	-5	-7	-8	-5	-3	5	-7	2	4	-7	-7	-5	-4	6	-3
Q	-6	-7	-8	-7	-4	0	-7	-8	-8	0	1	-6	-1	-8	8	-6	-5	-9	-8	-6
L	-1	-6	-7	-7	-4	-6	-7	-7	-7	5	4	-6	2	-1	-4	-6	-2	-6	-5	3
A	3	-6	-1	-5	-1	-6	-6	-4	-1	-4	-2	-6	3	7	-7	-1	-3	-4	-1	-4
H	0	-5	-6	-7	2	-3	-6	-6	3	2	4	-6	1	1	-6	-5	-4	6	-3	2
Q	-4	1	4	5	-7	4	-1	-4	-1	-6	-7	2	-6	-7	-3	-1	-4	- 7	-4	-6
L	-3	-6	-7	-7	1	-6	-7	-7	-7	3	5	-6	1	-2	-3	-4	-2	-6	-5	2
K	-1	3	1	0	-3	2	3	-4	1	-5	-5	4	-5	-6	-5	0	1	-6	-5	-4
D	-3	-2	4	3	-5	1	0	4	0	-7	-7	-1	-6	-7	-5	0	-2	-7	-4	-6
K	-3	5	-3	-4	-2	2	-2	-6	-2	-4	0	6	-1	-5	-6	-4	-5	-7	-4	-6
R	-2	3	2	-1	-5	0	-1	-4	4	-2	-3	2	-3	-6	3	-1	0	-6	-4	-1

M	-7	-11	-6	-14	-17	-9	-14	-19	-6	-12	-12	-8	20	-12	-11	-2	-6	-16	-4	-8
V	-13	-17	-13	-22	-23	-14	-21	-27	-11	-15	-15	-15	22	-8	-18	-9	-11	-20	2	-11
K	-19	-24	-21	-29	-27	-14	-28	-35	-19	-15	-15	-21	21	-16	-10	-15	-16	-29	-6	-17
S	-20	-30	-28	-36	-31	-20	-35	-42	-26	-10	-10	-27	23	-17	-14	-21	-18	-35	-11	-14
L	-13	-32	-24	-35	-28	-23	-36	-40	-22	-14	-14	-29	16	-7	-15	-18	-17	-35	-8	-17
										……………										
⋮																				
I	-20	-52	-40	-50	-22	-50	-53	-8	-44	-16	4	-49	-36	-39	-52	-34	-30	-44	-49	-5
G	-18	-45	-31	-40	-36	-43	-43	-6	-45	-23	-3	-40	-37	-41	-43	-30	-29	-45	-47	-14
G	-17	-44	-28	-39	-39	-43	-39	-3	-45	-18	2	-40	-38	-41	-40	-29	-28	-45	-48	-10

M	-107	-92	-90	-11	-28	-32	-38	-25	-28	-53
V	-145	-153	-147	-18	-49	-41	-37	-43	-43	-86
K	-214	-216	-189	-10	-69	-47	-70	-64	-57	-121
S	-238	-276	-224	-14	-83	-34	-89	-83	-71	-154
L	-228	-271	-207	-15	-71	-45	-72	-83	-71	-154
⋮										
D	-259	-114	-146	-26	-59	-54	-112	-6	-27	-33
…										
C	-201	-129	-87	-14	-28	-35	-101	-33	-22	-55
K	-188	-188	-115	-21	-42	-11	-99	-54	-40	-94

Fig. 5 Procedure of preparing feature sets for the predictor.

© 2013 Information Processing Society of Japan

where j represents each of the 20 amino acids. Each feature is calculated as $F_{i,p} = \sum_{j\in p} E_{i,j}$ ($i = 1, ..., m$, $j \in p$ means that j has the characteristic p). Finally, each value in the condensed and smoothed PSSM matrix is scaled to the range of $[-1, 1]$ according to a certain ratio. The procedure of preparing feature sets for the predictor is shown in Fig. 5.

2.10 Support Vector Machines (SVM) and 5-fold Cross-validation

Many studies [7], [8], [9], [10], [11], [12], [13], [14], [16], [17], [18], [19] have shown that SVM is powerful in dealing with high dimensional data, especially for binary classification. Identification of binding can be addressed as a two-classification problem, i.e., determining whether a given residue is a binding residue or not. In this study, the prediction model is trained by the libSVM software package which is written by in Chih-Jen [24]. Here, the Radial Basis Function (RBF kernel) is selected as the kernel function since RBF has been shown to be an optimal kernel in many cases. Both the capacity parameter c and kernel width parameter g are then optimized using a grid search approach [24]. 5-fold cross-validation is used to evaluate the performance of the developed methods, that is, the patterns are randomly divided into five sets. Four sets are used for training and the remaining one is used for testing, and the process is repeated until each set is used once for testing.

2.11 Four Prediction Methods

In order to analyze the impact on prediction of different feature-encoding schemes, four predictors using different features as input were designed, namely, SCPSSMpred (our new method), SVM$_{\_PSSM}$ [7], [8], [9], [10], [11], and our previous methods SVM$_{\_C\_PSSM}$ [17] and SVM$_{\_PSSM\_Physi}$ [18].

To develop the four classifiers, each method requires an outside sliding-window size, which indicates the length of the flanking regions considered affecting a central residue, and which will ultimately determine the dimensions of feature vectors in each method. Here, in order to facilitate a fair comparison with other studies, we chose the same outside-sliding window size as previous studies [8], [9], [10], [11], [17], [18], i.e., 17 for all four classifiers. In addition, SCPSSMpred requires a smoothing-window size, which is used for incorporating the information of flanking regions of a central residue. Detailed description for optimizing the smoothing-window size is explained later.

Predictor SCPSSMpred: this method adopts the smoothed and condensed PSSM as input. Each residue is encoded as a feature vector with 17 dimensions, i.e., (the residue to be predicted + 16 neighbors) × (10 physicochemical features).

Predictor SVM$_{\_C\_PSSM}$: this method adopts the condensed PSSM as input. Each residue is encoded as a feature vector with 17 dimensions, i.e., (the residue to be predicted + 16 neighbors) × (10 physicochemical features).

Predictor SVM$_{\_PSSM}$: this method adopts the standard PSSM as input. Each residue is encoded as a feature vector with 17 dimensions, i.e., (the residue to be predicted + 16 neighbors) × (20 amino acids).

Predictor SVM$_{\_PSSM\_Physi}$: this method adopts the standard PSSM and ten kinds of physicochemical properties of residues as input. Each residue is encoded as a feature vector with 17 dimensions, i.e., (the residue to be predicted + 16 neighbors) × (20 amino acids + 10 physicochemical features).

Table 1 Summary of the feature types and number of vector dimensions.

Method	Feature types	Dimensionality
SCPSSMpred (our new method)	The smoothed and condensed PSSM	$17 \times 10 = 170$
SVM$_{\_C\_PSSM}$ (our previous method [20])	The condensed PSSM	$17 \times 10 = 170$
SVM$_{\_PSSM}$ (P.S. Raghave's method [7-11])	PSSM	$17 \times 20 = 340$
SVM$_{\_PSSM\_Physi}$ (our previous method [19])	Physicochemical features and PSSM	$17 \times 30 = 510$

Table 1 summarizes the feature types and the number of vector dimensions of the four methods.

2.12 Evaluation Criteria

We adopted the evaluation criteria in CASP10 [25], which evaluates the performance of classifiers without bias using three indicators: the area under the corresponding ROC curve (AUC), ACC, and MCC (note that a method which achieves a higher ACC not always achieve a higher MCC [25]). The ROC plots with the AUC values were created using the R statistical package [26]. The sensitivity, specificity, true positive rate (TPR), false positive rate (FPR), accuracy, ACC, and MCC are defined as follows:

$$Specifity=\frac{TN}{TN+FP} \tag{1}$$

$$TPR=Sensitivity=\frac{TP}{TP+FN} \tag{2}$$

$$FPR=1-Specificity=\frac{TN}{TN+FP} \tag{3}$$

$$Accuracy=\frac{TP+TN}{TP+TN+FP+FN} \tag{4}$$

$$ACC=\frac{1}{2}(Sensitivity+Specificity) \tag{5}$$

$$MCC=\frac{(TP\times TN)-(FP\times FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{6}$$

where TP, TN, FP, and FN represent the number of true positives, true negatives, false positives, and false negatives respectively.

3. Results and Discussion

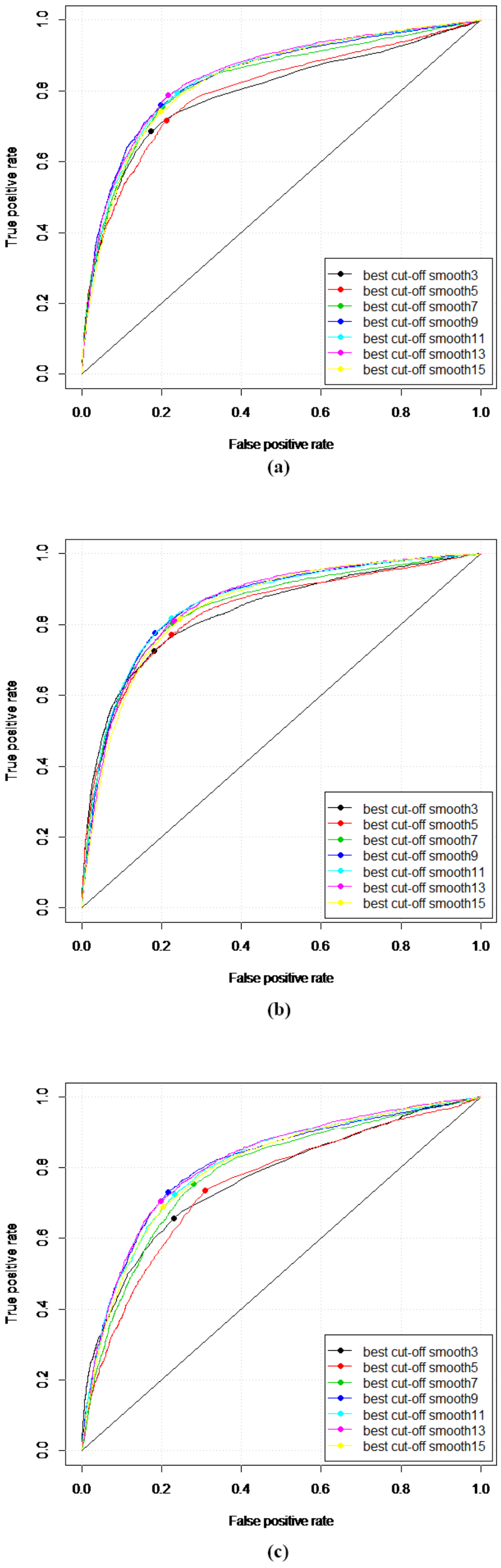
3.1 Performance with Different Smoothing-window Size

Using an outside sliding-window size of 17, the SCPSSMpred method was tested with different smoothing-window sizes, and the best ROC plot was chosen to represent the performance of each method. ROC plots of the SCPSSMpred method applied to NAD204, FAD191 and ATP200 with different smoothing-window sizes are shown in **Fig. 6** (a-c), and the respective optimal plots are shown in **Fig. 7** (a-c) separately.

3.2 Performance Comparison with Other Methods on Residue Level

After optimal ROC plots were obtained, the SCPSSMpred method was compared with the other three methods

© 2013 Information Processing Society of Japan

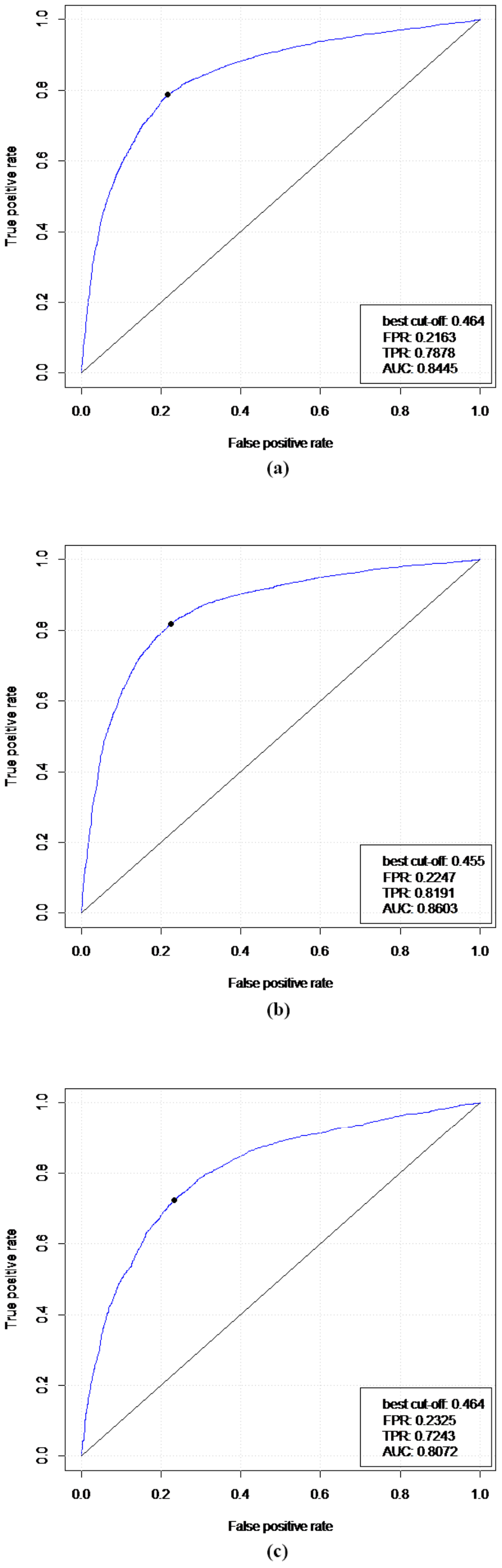


(a)

(b)

(c)

Fig. 6 **(a-c)** ROC of SCPSSMpred tested on NAD204 (a), FAD191 (b) and ATP200 (c). The sliding-window size was 17, and different smoothing-window sizes from 3 to 15 were tested. Optimal ROC plots for NAD204, FAD191, and ATP200 were obtained with smoothing-window sizes of 13, 11, and 13 respectively.



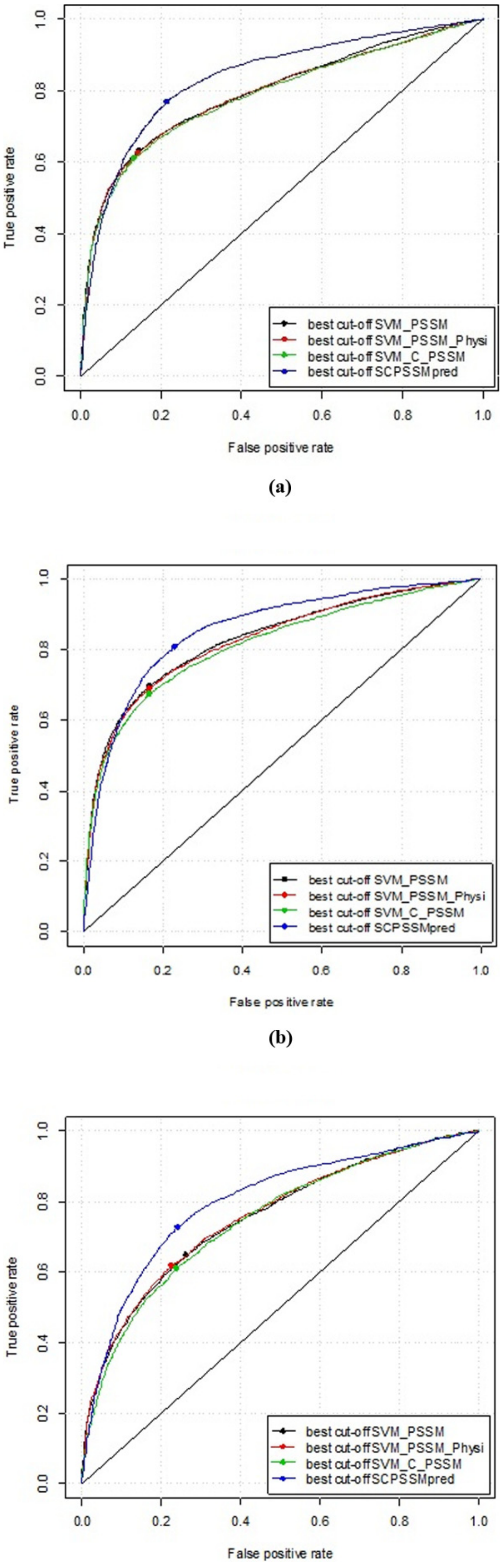
(a)

(b)

(c)

Fig. 7 (a-c) The best ROC plots of SCPSSMpred tested on NAD204 (a), FAD191 (b) and ATP200 (c).

© 2013 Information Processing Society of Japan



(a)

(b)

(c)

Fig. 8 (a-c) ROCs of different methods on the NAD204 (a), FAD191 (b), and ATP200 (c).

Table 2 Comprehensive performance of the four prediction methods on the three datasets (* is our proposed method).

Dataset	Method	ACC	TPR	FPR	MCC	AUC
NAD204	SCPSSMpred*	0.786	0.788	0.216	0.578	0.845
	SVM$_{\_C\_PSSM}$	0.746	0.618	0.127	0.507	0.787
	SVM$_{\_PSSM}$	0.742	0.621	0.138	0.497	0.796
	SVM$_{\_PSSM\text{-}Physi}$	0.742	0.621	0.137	0.499	0.791
FAD191	SCPSSMpred *	0.797	0.819	0.225	0.597	0.86
	SVM$_{\_C\_PSSM}$	0.767	0.691	0.157	0.53	0.815
	SVM$_{\_PSSM}$	0.762	0.708	0.184	0.528	0.83
	SVM$_{\_PSSM\text{-}Physi}$	0.763	0.684	0.159	0.531	0.827
ATP200	SCPSSMpred *	0.746	0.724	0.233	0.476	0.807
	SVM$_{\_C\_PSSM}$	0.687	0.641	0.268	0.375	0.748
	SVM$_{\_PSSM}$	0.686	0.69	0.318	0.372	0.754
	SVM$_{\_PSSM\text{-}Physi}$	0.689	0.619	0.241	0.382	0.757

(SVM$_{\_C\_PSSM}$, SVM$_{\_PSSM}$, and SVM$_{\_PSSM\_Physi}$). ROC plots of the four methods tested on NAD204, FAD191, and ATP200 are shown in **Fig. 8** (a-c), and a comprehensive performance of the four prediction methods tested on the three datasets is shown in **Table 2**.

Figure 8 and Table 2 demonstrate that the smoothed and condensed PSSM based predictor (SCPSSMpred) outperformed the other methods. Not only did it achieve the best ROC plots on the three datasets, but it also had the best ACC and MCC according to the evaluation criteria in CASP10.

From Fig. 8, it can also be seen that SVM$_{\_C\_PSSM}$, SVM$_{\_PSSM}$, and SVM$_{\_PSSM\_Physi}$ had similar performances. Among the three methods, SVM$_{\_C\_PSSM}$ used ten physicochemical properties to condense the PSSM, while SVM$_{\_PSSM}$ and SVM$_{\_PSSM\_Physi}$ used the traditional PSSM for prediction, that is, SVM$_{\_C\_PSSM}$ is 10-dimensional, SVM$_{\_PSSM}$ is 20-dimensional, and SVM$_{\_PSSM\_Physi}$ is 30-dimensional. This result illustrates that our condensing method can effectively reduce the redundant features in PSSM. It also indicates that binding is largely affected by the physicochemical properties of amino acids.

When comparing SCPSSMpred with SVM$_{\_C\_PSSM}$, it is clear that SCPSSMpred significantly outperformed SVM$_{\_C\_PSSM}$ method, because, unlike SVM$_{\_C\_PSSM}$, SCPSSMpred used the smoothing procedure. Thus, considering the impact of the neighboring residues on a central residue may greatly improve the prediction performance. It is worth noting that the PSSM itself is also affected by the limitation of the PSI-BLAST algorithm [22]. The SCPSSMpred method, which integrated ten physicochemical properties of residues, as well as information of neighboring residues, can reduce the complete dependence on the PSSM to some extent, because proteins that contain only a few homologous protein sequences exist, methods depending solely on PSSM for prediction may be ineffective.

3.3 Performance Comparison on Protein Sequence Level

A previous study [27] has shown that the performance of classifiers tested at a residue level may clearly be different from that at the protein sequence level. In practical applications, the number of non-binding sites is far larger than the number of binding sites in protein sequences. Therefore, for further analyzing the effectiveness of our method at the protein sequence level, the original 204 NAD-binding chains were analyzed as an example to test the performance of four methods. However, these chains may not be good test data, because some of their residues have been used for training the prediction model. Here, the best way to test at the sequence level is the leave-one-out method; however, due to the computational complexity, it is difficult to implement on the 204 sequences. Since the residues used for training prediction model are only 7.29% of the residues in the 204 NAD-binding sequences, these sequences can be used to test the relative effectiveness of the four methods on protein sequence level. Each sequence of the 204 proteins was therefore treated as a separate test set and used only once for testing individually. The ROC plot of the four methods tested on the 204 NAD-binding sequences is shown in **Fig. 9**, and the ACCs of the four methods at different thresholds are shown in **Fig. 10**. The accuracy distribution of the four methods tested on the 204 sequences is shown in **Fig. 11**.

Both Figs. 9 and 10 demonstrate that the SCPSSMpred method yielded a better ROC plot and a higher ACC than the other three methods. Figure 11 shows that most of the 204 sequences were more accurate when based on SCPSSMpred method rather than the other three methods. These results illustrate that the SCPSSMpred is also significantly better than the SVM$_{\_C\_PSSM}$, SVM$_{\_PSSM}$, and SVM$_{\_PSSM\_Physi}$ methods even when tested on the protein sequence level. From Fig. 11, it can also be seen that the SCPSSMpred and SVM$_{\_C\_PSSM}$ methods, which condensed the standard PSSM, performed better than the SVM$_{\_PSSM}$ and SVM$_{\_PSSM\_Physi}$ methods, which did not condense the PSSM. These data illustrate how low-dimensional features can bring better adaptability in dealing with imbalanced data, which meets the requirement for practical applications.

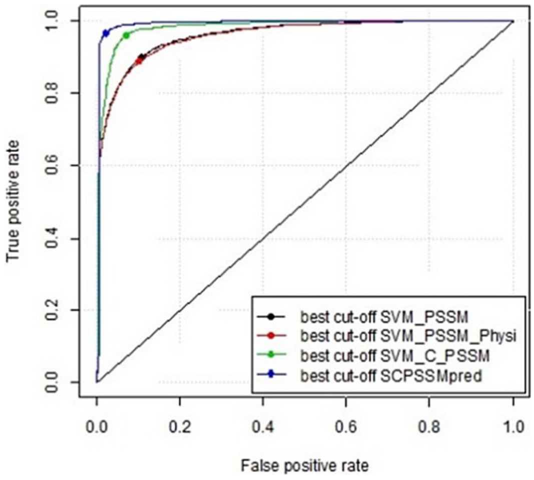


Fig. 9 ROC plots of the four methods tested on protein sequence level.

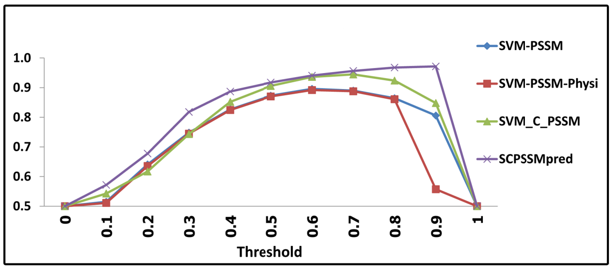


Fig. 10 ACC of the four methods at different thresholds.

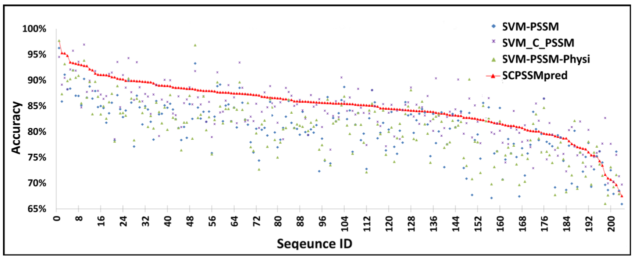


Fig. 11 Accuracy distributions of 204 NAD-binding sequences.

4. Conclusions

In this paper, we proposed a simplified PSSM encoding scheme containing only ten dimensions to replace traditional PSSMs for ligand-binding prediction. Although the simplified PSSM has only one-half of the feature dimensions of traditional PSSMs, it combines many features of traditional methods such as neighboring residue information, physicochemical properties, and evolutionary information. Comparing SCPSSMpred with three other traditional methods on three kinds of ligands (NAD, FAD, and ATP) showed that SCPSSMpred outperformed other methods without combining information of neighboring residues (SVM$_{\_C\_PSSM}$), using only the direct output of PSSM (SVM$_{\_PSSM}$) or using a simple connection of PSSM with ten physicochemical properties (SVM$_{\_PSSM\_Physi}$). Moreover, when tested at the protein sequence level, our method showed greater adaptability than other methods in dealing with unbalanced data.

Our study identifies several advantages of the simplified PSSM-based method. Firstly, it reiterates that ligand-binding is related to the arrangement of neighboring residues, and that the smoothing PSSM encoding scheme is effective for incorporating information of neighboring residues. Secondly, using the physicochemical properties of amino acids to condense the PSSM can largely reduce its redundant features. Thirdly, the SCPSSMpred method, which combines ten physicochemical properties of residues, can alleviate the complete dependence on PSSMs to some extent. For example, the SVM$_{\_PSSM}$ method is deeply dependent on evolutional information (PSSM), and because proteins having few homologous sequences exist, using only PSSMs for prediction may be ineffective in such cases. Lastly, considering the limited number of available samples, it is desirable that SCPSSMpred reduce the requirement for a large number of samples, which is a characteristic of high-dimensional feature space in machine learning. Moreover, it also can reduce the impact of over-fitting to noise data.

In summary, this paper not only demonstrates the necessity and importance of reducing redundant features in PSSM, but also reveals some hallmarks of nucleotide binding. A free Web server has been developed, which allows users to identify NAD-binding site in a given sequence using the model trained on our data set (http://webapp.yama.info.waseda.ac.jp/fang/ligand2.php).

References

[1] Ofran, Y., Mysore, V. and Rost, B.: Prediction of DNA-binding residues from sequence, *Bioinformatics*, Vol.23, No.13, pp.347–353 (2007).

[2] Moreira, I.S., Fernandes, P.A. and Ramos, M.J.: Hot spotsa review of the protein-protein interface determinant amino-acid residues, *Proteins: Structure, Function, and Bioinformatics*, Vol.68, No.4, pp.803–

© 2013 Information Processing Society of Japan

812 (2007).

[3] Chen, W.X. and Jeong J.C.: Sequence-based prediction of protein interaction sites with an integrative method, *Bioinformatics*, Vol.25, No.5, pp.585–591 (2009).

[4] Ma, B., Elkayam, T., Wolfson, H. and Nussinov, R.: Protein-protein interactions: Structurally conserved residues distinguish between binding sites and exposed protein surfaces, *Proc. National Academy of Sciences of the United States of America*, Vol.100, No.10, pp.5772–5777 (2003).

[5] Espadaler, J., Romero-Isart, O., Jackson, M.R. and Oliva, B.: Prediction of proteinprotein interactions using distant conservation of sequence patterns and structure relationships, *Bioinformatics*, Vol.21, No.16, pp.3360–3368 (2005).

[6] Dym, O. and Eisenberg, D.: Sequence-structure analysis of FAD-containing proteins, *Protein Science*, Vol.10, No.9, pp.1712–1728 (2001).

[7] Ansari, H.R. and Raghava, G.PS.: Identification of NAD interacting residues in proteins, *BMC Bioinformatics*, Vol.11, No.160, pp.1–8 (2010).

[8] Mishra, N.K. and Raghava, G.PS.: Prediction of FAD interacting residues in a protein from its primary sequence using evolutionary information, *BMC Bioinformatics*, Vol.11, No.S48, pp.1–6 (2010).

[9] Chauhan, J.S., Mishra, N.K. and Raghava, G.PS.: Prediction of GTP interacting residues, dipeptides and tripeptides in a protein from its evolutionary information, *BMC Bioinformatics*, Vol.11, No.301, pp.1–9 (2010).

[10] Kumar, M., Gromiha, M.M. and Raghava, G.PS.: Prediction of RNA binding sites in a protein using SVM and PSSM profile, *Proteins: Structure, Function, and Bioinformatics*, Vol.71, pp.189–194 (2007).

[11] Chauhan, J.S., Mishra, N.K. and Raghava, G.PS.: Identification of ATP binding residues of a protein from its primary sequence, *BMC Bioinformatics*, Vol.10, No.434, pp.1–9 (2009).

[12] Capra, J.A. and Singh, M.: Predicting functionally important residues from sequence conservation, *Bioinformatics*, Vol.23, No.15, pp.1875–1882 (2007).

[13] Terribilini, M., Lee, J.H., Yan, C., Jernigan, R.L., Honavar, V. and Dobbs, D.: Prediction of RNA binding sites in proteins from amino acid sequence, *RNA*, Vol.12, No.8, pp.1450–1462 (2006).

[14] Chen, K., Mizianty, M.J. and Kurgan, L.: Prediction and analysis of nucleotide-binding residues using sequence and sequence-derived structural descriptors, *Bioinformatics*, Vol.28, No.3, pp.331–341 (2012).

[15] Gonzalez, R.C. and Woods, R.E.: *Digital Image Processing*, Prentice Hall (2002).

[16] Cheng, W.C., Su, E.CY., Hwang, JK., Sung, TY., and Hsu, WL.: Predicting RNA-binding sites of proteins using support vector machines and evolutionary information, *BMC Bioinformatics*, Vol.9 (Suppl12), No.S6, pp.1–19 (2008).

[17] Fang, C., Noguchi, T. and Yamana, H.: Prediction of FAD binding residues with combined features from primary sequence, *International Proceedings of Computer Science and Information Technology*, Vol.34, pp.147–153 (2012).

[18] Fang, C., Noguchi, T. and Yamana, H.: Ligand-binding prediction based on a condensed position-specific scoring matrix, *International Journal of Data Mining and Bioinformatics*, In assessing.

[19] Shimizu, K., Hirose, S. and Noguchi, T.: POODLE-S: Web application for predicting protein disorder by using physicochemical features and reduced amino acid set of a position-specific scoring matrix, *Bioinformatics* , Vol.23, No.17, pp.2337–2338 (2007).

[20] Bauer, R.A., Günther, S., Jansen, D., Heeger, C., Thaben, P.F. and Preissner, R.: SuperSite: Dictionary of metabolite and drug binding sites in proteins, *Nucleic Acids Res.*, Vol.37, pp.195–200 (2009).

[21] Sobolev, V., Sorokine, A., Prilusky, J., Abola, E.E. and Edelman, M.: Automated analysis of interatomic contacts in proteins, *Bioinformatics*, Vol.15, pp.327–332 (1999).

[22] Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J.: Gapped BLAST and PSI-BLAST: A new generation of protein database search programs, *Nucleic Acids Res*, Vol.25, No.17, pp.3389–3402 (1997).

[23] NR, available from ⟨ftp://ftp.ncbi.nih.gov/blast/db/fasta/nr.gz⟩ (accessed 2013-01-05).

[24] Chang, C.C. and Lin, C.J.: LIBSVM: A library for support vector machines, *ACM Trans. Intelligent Systems and Technology*, Vol.2, No.27, pp.1–27 (2011).

[25] CASP10, available from ⟨http://predictioncenter.org/casp10/doc/presentations/CASP10_DR_KF.pdf⟩ (accessed 2013-02-26).

[26] R statistical software, available from ⟨http://www.r-project.org/⟩ (accessed 2013-02-26).

[27] Walia, R.R., Caragea, C., Lewis, B.A., Towfic, F., Terribilini, M., El-Manzalawy, Y., Dobbs, B. and Honavar, V.: Protein-RNA interface residue prediction using machine learning: An assessment of the state of the art, *BMC Bioinformatics*, Vol.13, No.89, pp.1–20 (2012).

Chun Fang was born in 1981. She is a Ph.D. student at Waseda University in Japan. She received her M.S. degree in Computer Science from Huazhong Normal University, Wuhan, China, in 2009. Her research interests include data mining, artificial intelligence and bioinformatics.

Tamotsu Noguchi is a principal research scientist of the Computational Biology Research Center (CBRC), AIST in Japan, from 2001. He is also a visiting professor of Research Institute of IT Biology and Mining of Waseda University from 2005. He received his Doctor of Engineering degree at Osaka University in 2001. His research interests include prediction of protein functions, prediction of secondary and tertiary structures in proteins, mechanism of protein folding and disorder regions in a protein.

Hayato Yamana received his Doctor of Engineering degree at Waseda University in 1993. He began his career at the Electro technical Laboratory (ETL) of the former Ministry of International Trade and Industry (MITI), and was seconded to MITI's Machinery and Information Industries Bureau for a year in 1996. He was subsequently appointed Associate Professor of Computer Science at Waseda University in 2000, and has been a professor in that department (as well as visiting professor at the National Institute of Informatics) since 2005. He is a member of IEEE and ACM.

(Communicated by *Kengo Sato*)

© 2013 Information Processing Society of Japan